**Is KOBE Going to Score His Next Shot: Let Us Tell You**

Allen Ansari, Yongjun Chu and [Solange Garcia de Alford](https://2ds.datascience.smu.edu/my/index.php?id=1394)

# **INTRODUCTION**

Kobe Bryant is one of the greatest basketball players in the history of National Basketball League. He scored 38,024 points over 20 years of his professional career and won 5 NBA championships. He took a lot of shots. Some went in the basket, while others didn’t. By identifying factors affecting his scoring potential, we may be able to predict or classify whether his next shot will get in the basket given the circumstances. The potential findings from this prediction may be interesting and beneficiary to those current or future NBA players who want to improve their scoring potential of a shot. Here, we attempted to develop a classification model that best describes the outcome of any shot he took in his career based on a publicly available dataset in which each shot was characterized by 29 different variables. Specifically, we wanted to build two different prediction models: a logistic regression model [1] and a Linear Discriminant Analysis (LDA) model [2] with cross-validation using a training partition to derive the classification rules of the model’s algorithms and a test partition of the data to apply the classification rules to predict the result of 5,000 shots Kobe made but with outcomes hidden from us.

**MATERIAL AND METHODS**

## **Data Description**

## The original data set contains a total of 30,697 shot attempts by Kobe Bryant in his 20 years’ career with related data to his shot attempts. The data was partitioned as follows:

## 25,697 records which would be used as the training set in the modeling.

## 5,000 records with the shot\_made\_flag values removed for use as test set in predictions.

The data contains 29 variables which are listed below including a brief description (**Table 1**). Some are continuous variables while others are categorical variables.

**Data Preprocessing**

During the initial modeling with logistic regression, we found that we needed re-coding of the explanatory categorical variable action\_type to avoid a quasi-complete separation issue [3]. This issue would cause the obtained parameter estimates and associated statistics of a model not reliable. After collapsing some of levels based on the order of hierarchical clustering, this issue was resolved. We re-leveled the categorical variable “matchup” to 2 levels (home and away) from original 74 levels. By working along with variable “opponent”, this recoded variable works just the same as the original one. We didn’t attempt to do any collinearity correcting as SAS (e.g. Proc Logistic) has the internal functionality to eliminate such issue. All the detail about how some variables were recoded can be found in the SAS code in the appendix. We found that transformations were not necessary for any of the variables during exploratory model fitting. We will address the outlier issue in the Results and Discussion section.

**Data Mining Approach and Evaluation**

For both logistic regression and LDA modeling, the probability that an event will occur is calculated, in other words, the probability that a shot will go into the basket. If the probability is over 50%, the outcome is predicted to be positive, otherwise it is negative. During the model evaluation stage, Cross Validation (CV) with one observation left out was applied with SAS to reduce potential biases. To further evaluate if obtained model sufficiently fits the data, we randomly partitioned the input dataset into two parts, training set (80% of total observations) and test set (20%). The default criterion was selected for each step of modeling evaluation for different techniques. For example, significance level (SL) was used for SAS Proc Logistic regression.

**Table 1**. Descriptions for each variable in the input data set.

|  |  |
| --- | --- |
| **action\_type:** The type of shot attempted, such as jump shot, dunk, etc.**combined\_shot\_type:** Classifies the shots under 6 larger categories: Bank Shot, Dunk, Hook Shot, Jump Shot, Layup, and Tip Shot.**Matchup:** The two teams in the specific match. Since Kobe was always on the Lakers and opponent contains all the information in matchup, we decided to reduced number of levels for this variable by dividing all games into Home and Away category.**Opponent:** Opponent in the specific match.**Season:** The basketball season (2000, 2001, etc.)**shot\_type:** includes categories 2pt or 3pt.**shot\_zone\_area:** Area from which shot was attempted (Right, Left, Center, Back Court, Right Center, Left Center)**shot\_zone\_basic:** Further area information (Mid-range, restricted area, in the paint, above the break 3, backcourt, left corner 3, right corner 3)**shot\_zone\_range:** Range (<8 ft, 8-16, 16-24, 24+, backcourt)**team\_name:** Name of Kobe’s team, the Lakers, so we decided to remove it from dataset.**arena\_temp:** average temperature of are*na***attendance**: Number of people who watched the game | **avgnoisedb:** Average noise level in the arena in decibels**game\_date:** Date of the specific match.**game\_event\_id:** **game\_id:** NBA game ID**lat:** The latitude of Kobe’s position during the shot attempt.**loc\_x:** The x-location on the court.**loc\_y:** The Y-location on the court.**Lon:** The longitude of Kobe’s position during the shot attempt.**minutes\_remaining:** The minutes remaining in the specific match**period**: The period in the specific match**playoffs:** binary, 1/0 values**recId****seconds\_remaining:** The seconds remaining in the specific match**shot\_distance: The** distance from which the shot was attempted, in ft.**shot\_id:** (from 1 to 30,697) of the attempted shot**shot\_made\_flag:** it is response variable and indicates if shot was successful (as 1) or not (as 0)**team\_id**: ID of Kobe’s team. Always the Lakers, so removed |

**RESULTS AND DISCUSSIONS**

## **Exploratory Data Analysis**

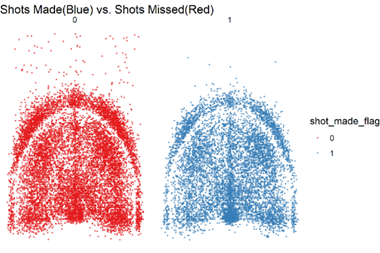
We explored some of predictors to find meaningful relationships with shot\_made\_flag, the dependent variable. First, we examined correlations between variables. **Figure 1A** shows there is extensive correlation between loc\_y and lat or shot\_distance, which make sense considering they are basically describing the same thing, the distance from basket for Kobe. Additionally, there is high correlation between attendance and average noise in decibels. Obviously, more people in an arena would generally produce more noise.



**Figure 1. The correlation matrix of 12 continuous variables.**

Next, we examined location data as shown in **Figure 2A**. By plotting latitude vs longitude a visualization of the shots Kobe made and missed by location is depicted. It is rather difficult, at first glance, to discern any differences. One point that becomes obvious, however, is the impact of range. There are more misses than makes at longer ranges, meaning the 3-point line and beyond. Within the 3-point area the data is too noisy to analyze. **Figure 2B** provides a visualization of shot\_zone\_area, showing the on-court representation of each zone. The accuracy of shots in each zone shown in **Figure 2C**. As expected, the accuracy for shots from the backcourt is extremely low.

1. (B) (C)

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Figure 2. The visualization of shot location on the court.** (A) The shot coordinates on the court. (B) The shot zone area on the court. (C) Shot making/missing percentage within each shot zone area.

We also examined the shot\_zone\_basic variable. **Figure 3A** provides a visualization of the on-court locations for each level of this variable. **Figure 3B** shows the shot accuracy for each type of location. Kobe’s accuracy by shot\_zone\_basic does not vary by that much, except in the back-court location where he scored poorly.

1. (B)

A close up of a piece of paper

Description automatically generatedA screenshot of a cell phone

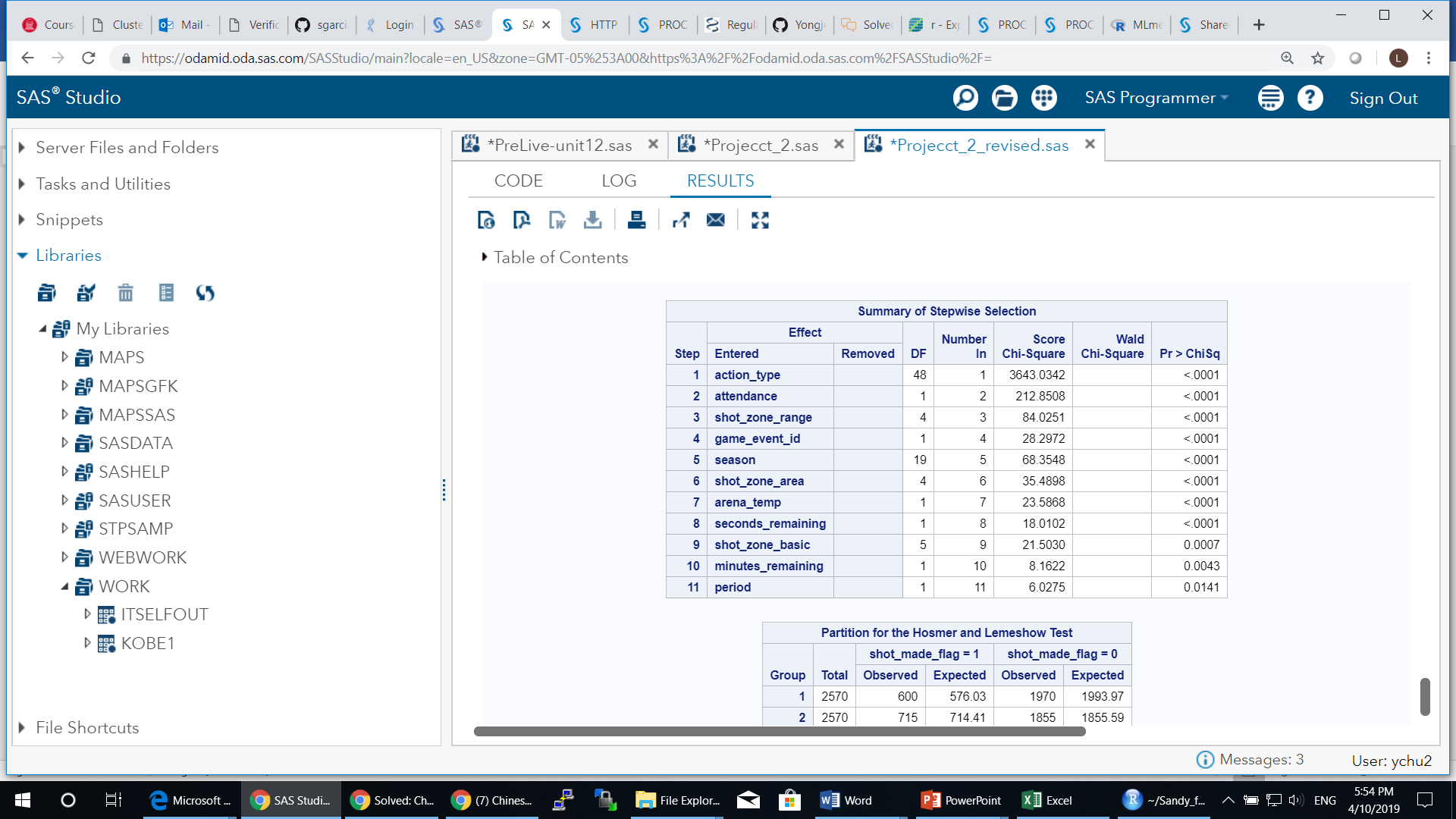
Description automatically generated

**Figure 3. The visualization of shot zone area on the court.** (A) The distribution of shot zone area on the court. (B) Shot making/missing percentage within each shot zone basic location.

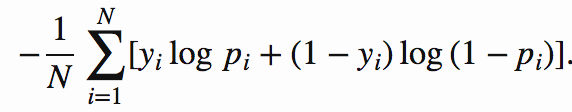
**Logistic Regression Model More Accurate at Predicting the outcome of a shot**

To best model if Kobe’s shot will go into the basket, two powerful classification methods were chosen: logistic regression analysis and linear discriminant analysis (LDA). To compare how these two models might perform, we first generated a train data set (80%) and test dataset (20%) based on the whole dataset containing 25697 observations. For logistic regression, we first identified 11 statistically significant predictor variables based on the training dataset: action\_type, attendance, shot\_zone\_range, game\_event\_id, season, shot\_zoen\_area, arena\_temp, seconds\_remaining, shot\_zone\_basic, minutes\_remaining and period (**Table 2**). Action\_type appears to be the most significant variable.

**Table 2**. The significant variables identified in Proc Logistic stepwise selection method.



These variables were found using Proc Logistic procedure with stepwise selection method. Forward selection method yielded the same set of significant variables. For LDA modeling, to satisfy its assumption, we only kept the quantitative variables in the modeling (SAS Proc Discrim). These quantitative variables are: shot\_distance, arena\_temp attendance, avgnoisedb, game\_event\_id, lat, loc\_x, loc\_y, lon, minutes\_remaining, period, seconds\_remaining. A normal distribution was assumed and the covariance matrix was pooled during modeling. The regression parameter estimates were then derived for each model and further were applied to the test dataset. We obtained the mis-classification rate, sensitivity, and specificity for each case with test dataset. To obtain the AUC (Area Under the Receiver Operating Characteristic (ROC) Curve), we used the R package, “MLmetrics”, which calculates the AUC based on the ROC between true positive rate and false positive rate. To assess the model fit, we also calculated the log loss function, which is defined as the following,



Where N is the total number classifications, yi is the shot\_made\_flag and pi is the probability from the model of each outcome (shot made). **Table 3** shows the comparison results.

**Table 3.** The comparison of model fitting with test dataset for Logistic regression model and LDA.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Logistic Regression Model** | **Linear Discriminant Analysis** |
| AUC | 0.6877 | 0.5680 |
| Mis-Calculation Rate | 0.3257 | 0.4276 |
| Sensitivity | 0.4613 | 0.6245 |
| Specificity | 0.8485 | 0.5290 |
| Log LOSS function value | 0.6182 | 0.6972 |

The comparison results in **Table 3** clearly suggested that logistic regression model outperformed the LDA with the input data. The former has higher AUC value by more than 20% than the latter. Logistic modeling also has better specificity and lower log LOSS function value. All these results indicate that logistic regression model fits the data clearly better. Therefore, we used the logistic regression model to further study specific questions of interest with the whole dataset. One likely reason that logistic regression worked better here is that logistic regression can take either type of variables, continuous or categorical while LDA only takes continuous or quantitative variables. One could convert some of categorical or binary variables into continuous variables with specific domain knowledge.

**Checking the Fitting of Logistic Model on Whole Input Data by Cross Validating (with one left out method)**

We applied logistic regression model to check how it performed on the whole input dataset (25697 observations). In the modeling, we first used stepwise or forward methods to identify significant variables. With no surprise, the 11 variables identified earlier with training dataset were identified again here. The Hosmer and Lemeshow Goodness-of-Fit test gave a p-value of 0.48, suggesting there is no significant evidence that the logistic modeling is a bad fit. To find out how well this model with 11 variables fit the data, we carried out the classification for each observation in the whole input dataset. To reduce the biases due to the predicting of itself, we used cross validation by leaving one out method, which is easily carried out in SAS Proc Logistic procedure. In this method, all the rest of observations are used like a training dataset and prediction is carried out on that one observation. The fit statistics is shown in **Table 4**. The fit results suggest that the logistic model with 11 variables fit the data well, with a log LOSS function value of 0.605. The ROC curve is shown in supplementary (**Figure S1**). When plotting the diagnostic results for finding inferential data points, we found that observation 5850 may have relatively large influence on model fit and parameter estimates (**Figure S2**). Further investigation suggests the overall model fit didn’t change much without this observation. Therefore, we kept this observation in the dataset.

**Table 4**. The modeling fit results of logistic regression with whole input dataset.

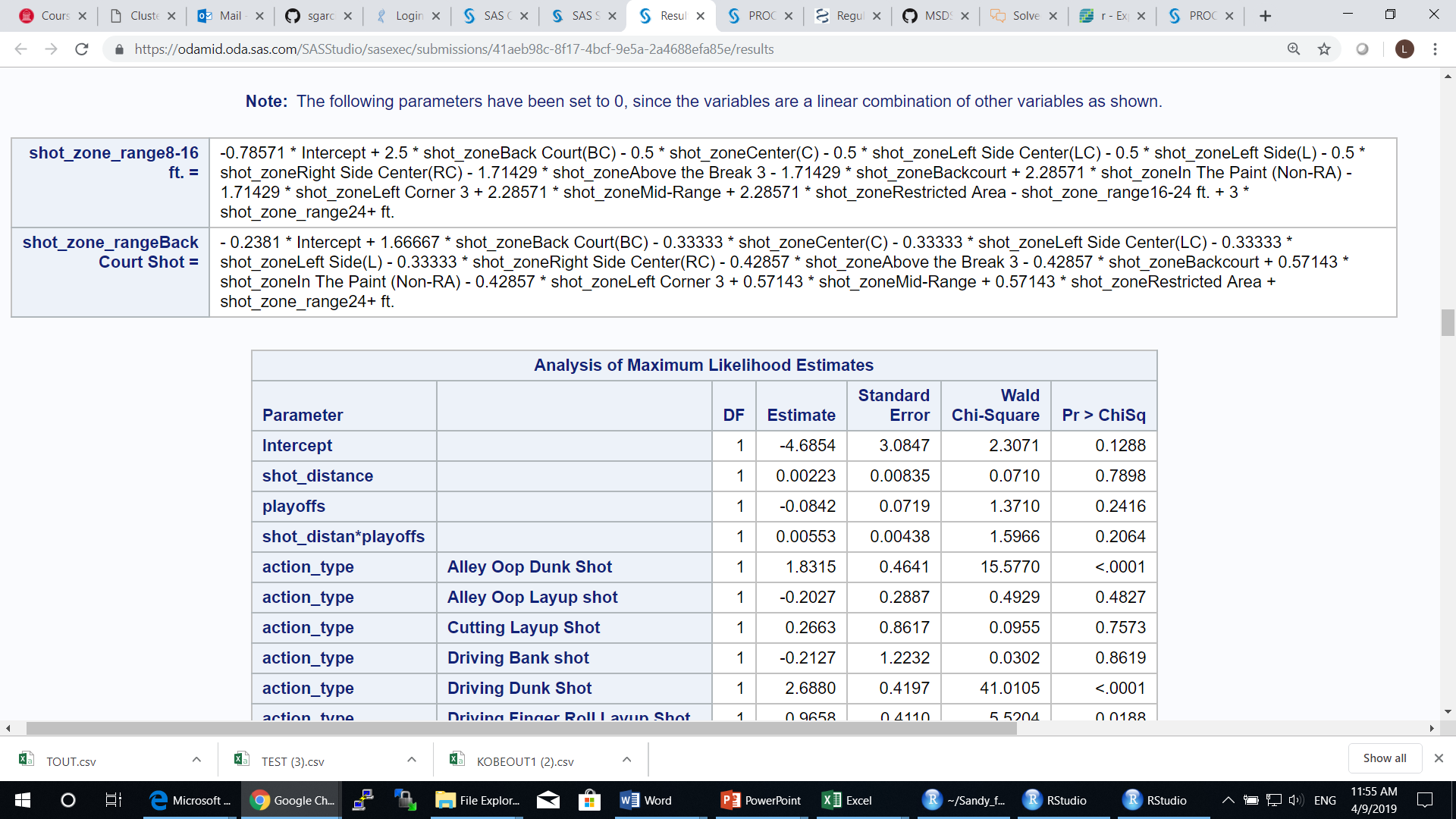
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AUC** | **Mis-Calculation Rate** | **Sensitivity** | **Specificity** | **Log LOSS function value** |
| 0.6956 | 0.316 | 0.4633 | 0.8618 | 0.605 |

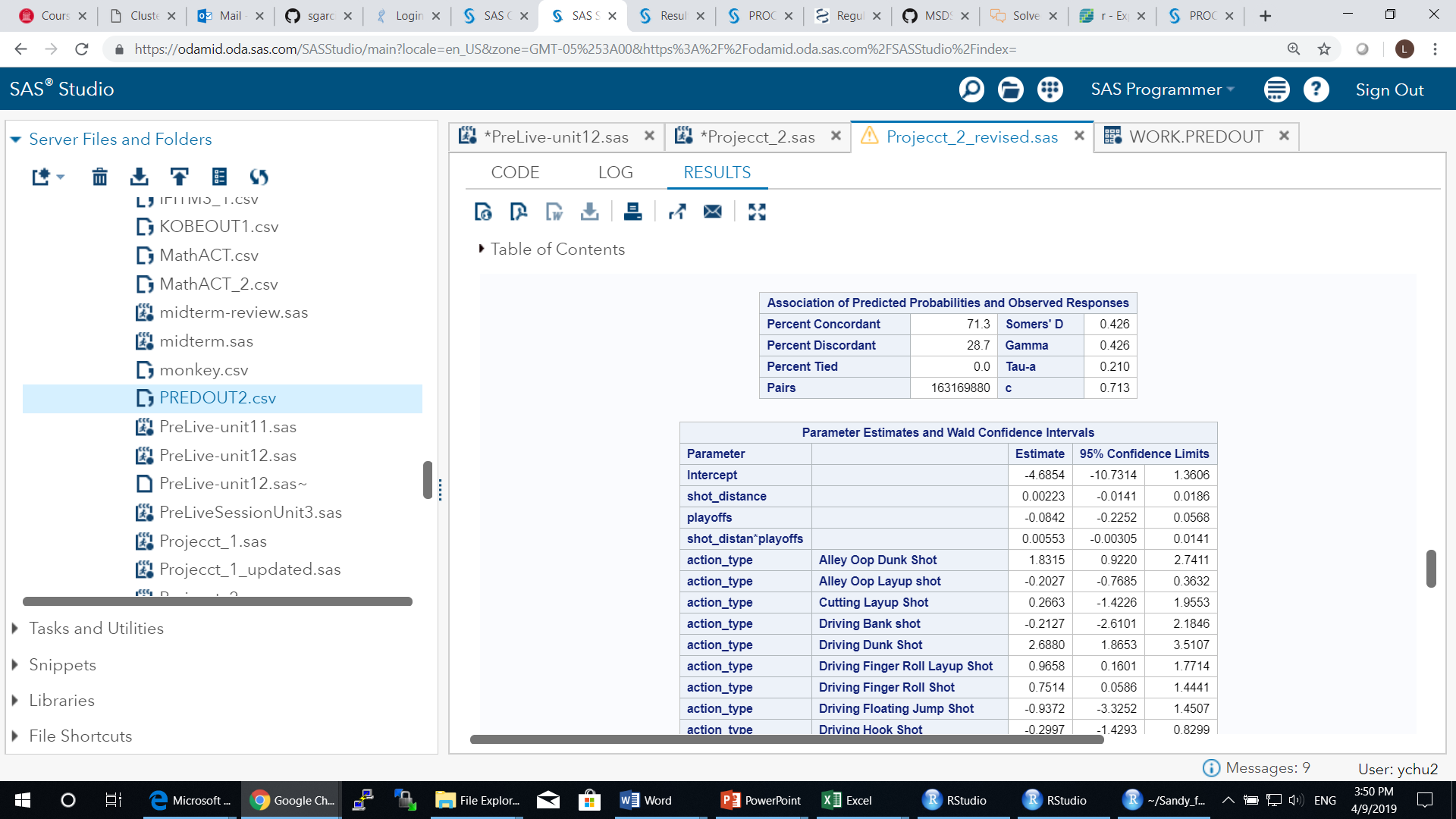
**Shot Distance and Playoffs not Significant Predictors on Making a Shot**

A common perception is when a basketball player is closer to the basket, it would be easier to make a shot. The reality is often more complicated than that. This is simply because the defensive players can do a better job of defending when an offensive is closer to the basket for scoring. Another interesting question is how Kobe performed in playoff games relative to the regular season games and whether the playoffs affect the relationship between shot distance and his scoring potential.

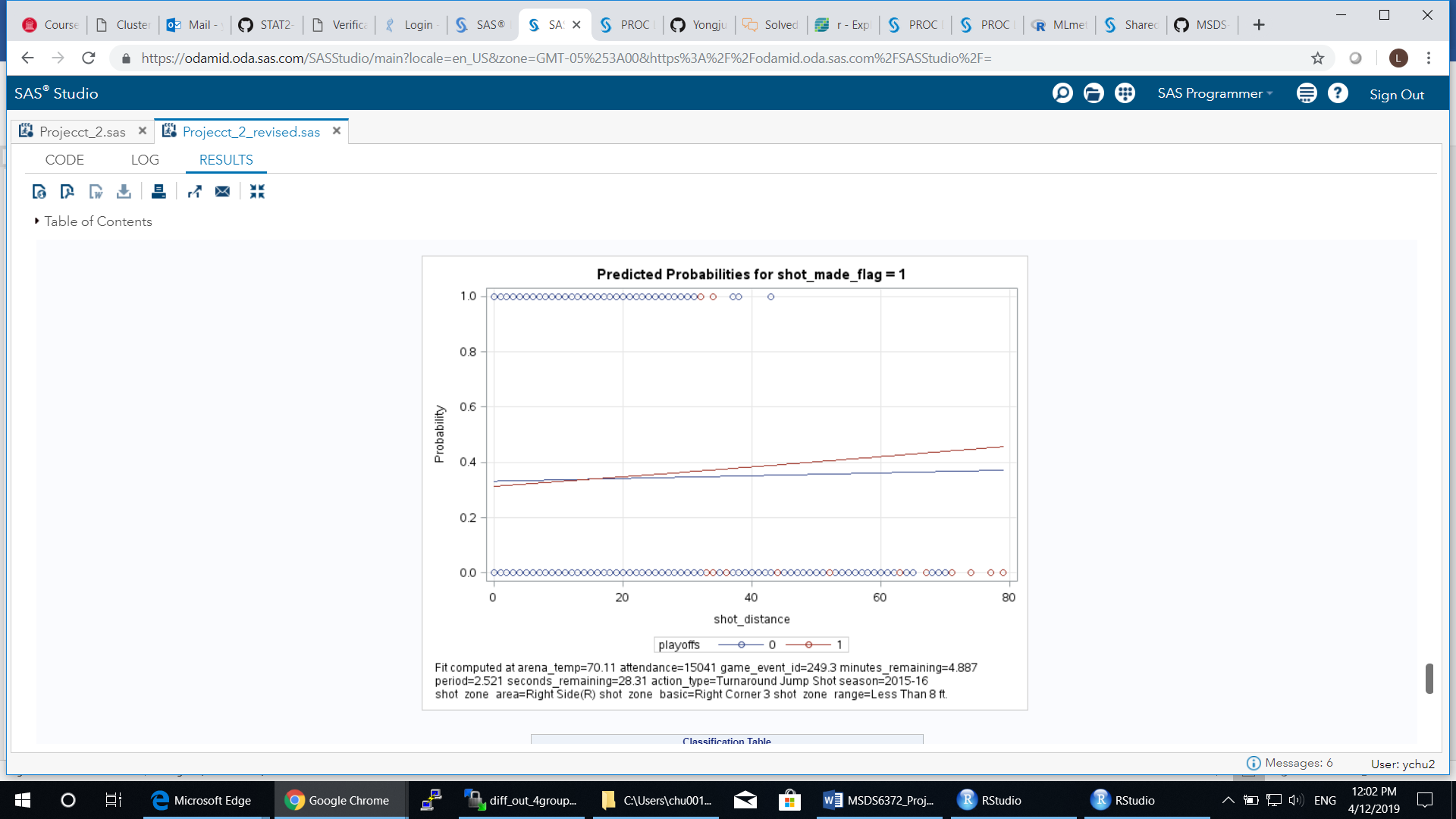
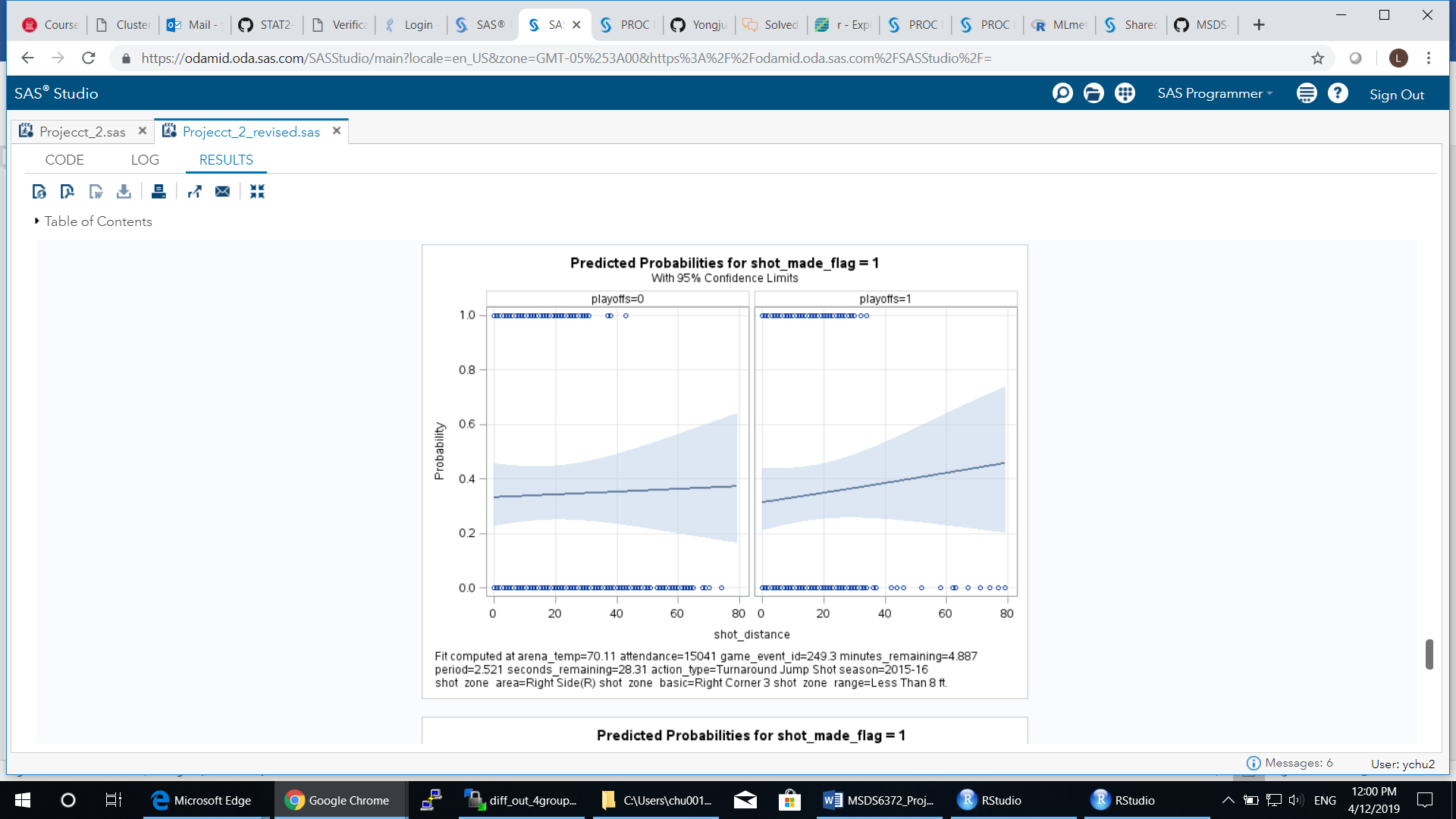
To answer these questions, we applied a logistic regression model to the whole input dataset (25697 observations). Although shot\_distance and playoffs were not found to significant variables, we included both and the interaction term, shot\_distance\*playoffs, along with other 11 significant variables in the model fitting. The Hosmer and Lemeshow Goodness-of-Fit test gave a p-value of 0.3678, suggesting there is no evidence that the logistic modeling with 14 variables is a bad fit. The fitting statistics of shot\_distance, playoffs, and their interaction term are shown in **Table 5**. Based on the fitting statistics, log(odds of making a shot) increases 0.00223 for every unit increase of shot\_distance, or the odds of scoring on a shot is 1.0022 when Kobe is one unit of distance away from the basket. Although this seems odd, it can be understood in this way: when he is not close to basket, he has more space to operate by using favorable moves to score. When he is closer to basket, pressure goes up as more defensive players would crowd him, making him harder to score. The relationship between the probabilty of Kobe making a shot and the distance he is from the hoop is shown in **Figure 4** after accounting for other variables in the model. It appears that there is a slightly positive linear relationship between distance and probalibity to score. The 95% CI band is also shown in the **Figure 4**.

**Table 5**. The fitting statistics of shot\_distance, playoffs, and shot\_distance\*playoffs.





On the other hand, the p-vlaue of 0.7898 and a 95% CI of [-10.73, 1.36], which contains zero for the coefficient estimate, suggest that there is really no significant relationship between shot\_distance and whether he will make shots after accounting for other variables. This indicates that other significant variables aleady explained well the chances that he will make a shot or not. Similar explainatin can be applied to playoffs, which has a p-value of 0.2416 and 95% CI [-0.2252, 0.0568]. The estimated coefficient for interaction term between shot\_distance and playoffs is not significant either (p-value=0.2064, 95% CI [-0.003, 0.014]) (**Table 5**). Based on this analysis, we concluded that whether Kobe made a shot or not is not significantly related to the distance he is from the hoop. The relationship between the distance Kobe is from the basket and the odds of him making the shot is not different if they are in the playoffs.



**Figure 4**. **The relationship between probability of Kobe scoring on a shot and his distance to the hoop under different levels of playoffs (1 or 0)**.

**Predicting the Outcomes of 5000 Kobe’s Shots**

By using the best model identified, that is, the logistic regression model, we predicted the outcome of 5000 shot Kobe took during his career. In the prediction, we used all the input data points of (25697 observations) as the training dataset and 11 significant variables we identified earlier. The prediction results have been uploaded.

**CONCLUSIONS**

In summary, we have found that logistic regression model is a better fit than LDA for the input data presented here. The logistic model yields higher prediction accuracy and sensitivity and lower log LOSS function value of Kobe’s shot outcome. Eleven variables were found to be significantly affecting Kobe’s scoring potential, particularly the action\_type, that is, the move to shoot the ball. Some moves are far easier to score than others. This finding may be interesting to those NBA players.

It may seem obvious that the closer a basketball player to the basket, the odds of scoring is increasing. This is expected if there is no defense presented. Our investigation of this question with Kobe’s shooting dataset suggests that this is not the case. Distance is not that important. Other factors affect his scoring potential more significant. For example, when Kobe is close to the basket, the action type of a slam dunk has much higher chance to score than a bank shot even he had the equal distances to the hoop at both situations. We also looked at if the playoffs changed his scoring potential with regard to his distance from basket. The results suggest no existing of such a significant impact from being in the playoff games.

**REFERENCES**

1. Agresti, A. (2002), Categorical Data Analysis, Second Edition, New York: John Wiley & Sons.

2. Rao, C. R. (1973), Linear Statistical Inference, New York: John Wiley & Sons.

3. Allison, Paul, D (2008), Convergence Failures in Logistic Regression, Philadelphia, PA, SAS Global Forum 2008, Paper 360-2008.